Detección de Patrones de Bajo Rendimiento Académico y Deserción Estudiantil con Técnicas de Minería de Datos

Ricardo Timarán Pereira, Ph.D.
Departamento de Sistemas, Facultad de Ingeniería, Universidad de Nariño
San Juan de Pasto, Nariño, Colombia
ritimar@udenar.edu.co

RESUMEN

En este artículo se presentan los resultados de la investigación realizada en la Universidad de Nariño (Colombia) cuyo objetivo fue determinar en la comunidad universitaria perfiles de bajo rendimiento académico y deserción estudiantil aplicando técnicas de descubrimiento de conocimiento, a partir de los datos almacenados en las bases de datos durante los últimos 15 años. Este proceso se apoyó con TariyKDD, una herramienta de minería de datos de distribución libre, desarrollada en los laboratorios de DCBD del Departamento de Ingeniería.

Palabras Claves: Detección de Patrones, Bajo Rendimiento, Deserción Estudiantil, Minería de Datos.

1 INTRODUCCIÓN

Debido al avance de la tecnología en los sistemas computacionales, se hace indispensable y necesaria la utilización de tecnologías informáticas que contribuyan a resolver ciertos problemas que sin la utilización de ellas, se haría prácticamente imposible el tratamiento de los mismos, brindando soluciones eficientes y sustentadas en la realidad para aplicarlas en el contexto en el que se encuentran. Una de estas tecnologías es la minería de datos, en la que se fundamentó todo el proceso investigativo de este proyecto.

La Universidad de Nariño es una institución pública de educación superior cuya área de influencia es el suroccidente de Colombia, cuya sede principal se encuentra en la ciudad de Pasto, capital del departamento de Nariño. En ella se encuentra la mayoría de estudiantes universitarios de la región. Los estudiantes de educación secundaria aspiran obtener un cupo en ésta, por su calidad educativa, y prestigio de sus egresados. Desafortunadamente, en algunos casos, cuando el estudiante se matricula a un determinado programa, su rendimiento no es el esperado, generando índices de deserción altos y bajo rendimiento académico. Por lo tanto se genera un interrogante acerca de cuáles son las causas que motivan la deserción y/o el bajo rendimiento y que perfiles tienen este tipo de estudiantes.

De acuerdo a datos obtenidos, actualmente se encuentran matriculados en los diferentes programas de pregrado 8136 estudiantes distribuidos en 11 facultades (ver tabla 1), de los cuales el 27.94% han perdido dos veces la misma asignatura, el 3.44% han perdido tres veces y el 0.19% cuatro veces. Así mismo el 6.71% tienen un promedio menor a 3, el 66.85% tienen un promedio entre 3 y 4 , y el 26.22% tienen un promedio mayor que 4 sobre 5. Además el promedio del índice de deserción está por encima del 32% en la última corte de los programas.

Tabla 1. Número de estudiantes por facultad.

Facultad	No. Estudiantes
Artes	902
Ciencias Agricolas	618
Ciencias de la Salud	243
Ciencias Económicas y Administrativas	1408
Ciencias Humanas	1341
Ciencias Naturales Y Matematicas	687
Ciencias Pecuarias	523
Derecho	510
Educación	435
Ingenieria	1188
Ingenieria Agroindustrial	281
Total	8136

En este artículo se describe el proceso de descubrimiento de patrones que determinen en la comunidad universitaria perfiles de bajo rendimiento académico y deserción estudiantil aplicando técnicas de minería de datos, a la información almacenada en las bases de datos durante los últimos 15 años, cuyos resultados permitan predecir a los posibles estudiantes, candidatos a encontrarse en estos estados y tomar los correctivos necesarios a tiempo que ayuden a minimizar estos factores y conlleven al mejoramiento de la calidad educativa en la universidad. Este proceso se apoyó con TariyKDD [11], una herramienta de minería de datos de distribución libre, desarrollada en los laboratorios de KDD del Departamento de Ingeniería de Sistemas de la Facultad de Ingeniería de la Universidad de Nariño.

El resto del artículo está organiza en secciones. En la sección 2, se abordan los conceptos preliminares que se tuvieron en cuenta para la ejecución de este proyecto. En la sección 3 se describe todo el proceso de descubrir patrones de bajo rendimiento académico y deserción estudiantil en la Universidad de Nariño. Finalmente, en la sección 4, se presentan las conclusiones y recomendaciones.

2. CONCEPTOS PRELIMINARES

2.1 Proceso de Descubrimiento de Conocimiento en Bases de Datos

El Descubrimiento de Conocimiento en Bases de Datos (DCBD) es básicamente un proceso automático en el que se combinan descubrimiento y análisis. El proceso consiste en extraer patrones en forma de reglas o funciones, a partir de los datos, para que el usuario los analice. Esta tarea implica generalmente preprocesar los datos, hacer minería de datos y presentar resultados [4][7][9]. Este proceso es interactivo e iterativo, involucra numerosos pasos con la intervención del usuario en la toma de muchas decisiones. En la figura 1 se muestran las etapas del proceso DCBD.

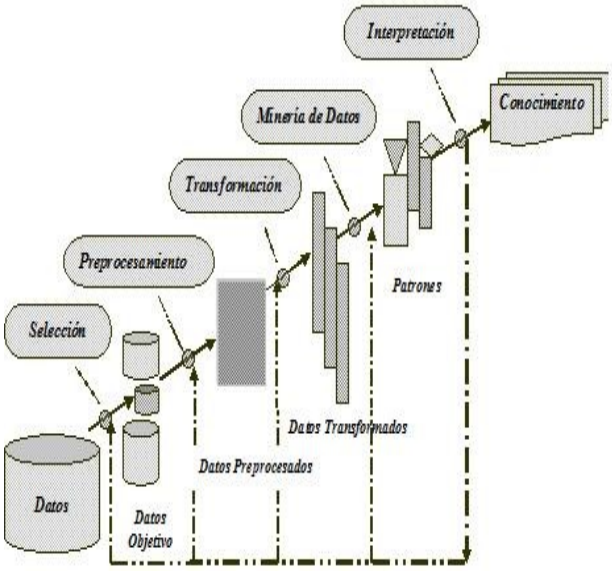

Figura 1. Etapas del proceso DCBD

2.1 Herramienta TariyKDD

TariyKDD, una herramienta para el Descubrimiento de Conocimiento, débilmente acoplada con un SGBD, desarrollada en el laboratorio KDD del departamento de Ingeniería de Sistemas de la Universidad de Nariño (Colombia), bajo software libre. Esta herramienta está compuesta por cuatro módulos: el módulo de *conexión* que permite la recuperación de datos desde archivos planos y bases de datos relacionales, el módulo de *utilidades* con clases y librerías comunes, el módulo *kernel* donde se encuentran los filtros que permiten realizar los procesos de limpieza y transformación de datos, los algoritmos de minería de datos para las tareas de Asociación y Clasificación y los programas de visualización de datos, y el módulo de *interfaz gráfica de usuario* que facilita la interacción del usuario con la herramienta de una manera amigable[11]. En *TariyKDD* se encuentran implementados los algoritmos *Apriori* [1], *FPGrowth* [6] *y EquipAsso* [12][13] para la tarea de Asociación y los algoritmos *C4.5* [10] *y Mate-tree* [14] para la tarea de Clasificación. La arquitectura de TariyKDD se muestra en la figura 2.

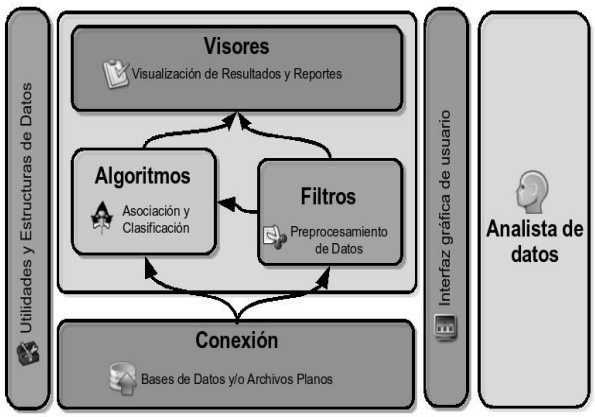

Figura 2. Arquitectura de la herramienta TariyKDD

3 PROCESO DE DETECCIÓN DE PATRONES DE BAJO RENDIMIENTO Y DESERCIÓN

3.1 Etapa de Selección

El objetivo de esta etapa es obtener las fuentes de datos internas y externas que sirven de base para el proceso de Minería de Datos. Como fuente interna, se seleccionó la base de datos histórica de los estudiantes de la Universidad de Nariño, compuesta por información personal y académica de 46173 estudiantes. Como fuente externa, se seleccionó la información de los colegios de educación secundaria del país, que se obtuvo con el Ministerio de Educación Nacional de Colombia.

Estas fuentes de datos se integraron en la base de datos UDENARDB, construida con el sistema gestor de base de datos PostgreSQL. UDENARDB la componen siete tablas, cuyas descripciones se pueden ver en la tabla 2.

3.2 Etapa de Preprocesamiento de Datos

El objetivo de esta etapa es obtener datos limpios, i.e. datos sin valores nulos o anómalos que permitan obtener patrones de calidad. . Por medio de consultas ad-hoc sobre la base de datos UDENARDB, se analizó minuciosamente la calidad de los datos contenidos en cada uno de los atributos de las tablas y se dejaron únicamente los atributos más relevantes para la investigación y aquellos que no contenían valores nulos. Como resultado de esta etapa, quedaron únicamente datos de 20329 estudiantes, para su posterior análisis. De la tabla alumnos se seleccionaron 19 atributos, de la tabla carreras 4, de la tabla facultades 2, de la tabla materias 3, de la tabla notas 8, de la tabla liquidación 12 y de la tabla colegios 3 atributos. Los atributos seleccionados de las diferentes tablas en su gran mayoría no contenían valores nulos ni anómalos (*outliers*), pero en aquellos casos que se presentaban, estos fueron reemplazados utilizando técnicas estadísticas tales como la media y la moda o derivando sus valores a través de otros como por ejemplo la edad de ingreso del estudiante conocida la fecha de ingreso y la fecha de nacimiento.

Tabla 2. Descripción de tablas de la base de datos UDENAR

Tablas	No. Atributos	Descripción
ALUMNOS	69	Se encuentran todos los datos personales del estudiante.
CARRERAS	10	Se encuentra información de todas las carreras existentes en la Universidad de Nariño
FACULTADES	4	Contiene información de las facultades de la Universidad de Nariño.
MATERIAS	4	Se encuentran toda la información de las materias existentes en el plan académico de cada carrera.
NOTAS	8	Contiene información de las notas por materia de cada estudiante.
LIQUIDACION	27	Se encuentra toda la información financiera del estudiante
COLEGIOS	7	Contiene información de los colegios del país

3.3 Etapa de Transformación de Datos

En la etapa de transformación, se buscan características útiles para representar los datos dependiendo de la meta del proceso de minería de datos. Se utilizan métodos de reducción de dimensiones o de transformación para disminuir el número efectivo de variables bajo consideración o para encontrar representaciones invariantes de los datos [5].

En esta etapa se construyó el conjunto de datos UDENAR.DAT, integrando los atributos de las diferentes tablas de la base de datos UDENARDB. Se eliminaron los atributos que eran llaves primarias de las tablas, se construyeron nuevos atributos (ver tabla 3) y se discretizaron los atributos continuos, es decir, se transformaron los valores numéricos en atributos discretos o nominales. Algunos de los atributos discretizados se muestran en las tablas 4,5 y 6.

Por otra parte, el conjunto de datos UDENAR.DAT se adecuó al formato ARFF (*Attribute Relation File Format*), utilizado por la herramienta TariyKDD para importar los datos.

La estructura del formato ARFF [15] es la siguiente:

- Cabecera: se define el nombre de la relación y su formato es el siguiente:
  @relation <nombre-de-la-relación>
- Declaraciones de los atributos. En esta sección se declaran los atributos que compondrán el archivo arff con su tipo. La sintaxis es la siguiente:
  @attribute <nombre-del-atributo> <tipo>
- Sección de datos. Se declaran los datos que componen la relación separando entre comas los atributos y con salto de líneas las relaciones.

Tabla 3. Descripción de nuevos atributos del conjunto de datos UDENAR.DAT

Atributo	Descripción
Ingresos	Establece un valor real actualizado de ingresos familiares del estudiante. Para ello relaciona los campos ingresos_familiares de la tabla alumnos e ingresos de la tabla liquidación.
Edad	Determina que edad tiene actualmente el estudiante; para ello se relacionaron los campos fecha nacimiento de la tabla alumnos y la fecha actual.
edad_ing	Establece la edad en la que ingreso el estudiante. Para crearlo se relaciono el campo fecha de ingreso y la fecha de nacimiento del estudiante.
val_matricula	Determina el valor real que paga el estudiante por concepto de matricula financiera; relaciona los valores de los campos nueva_matricula y de nuevos servicios de la tabla liquidación.
Claseal	Determina que estudiantes han reingresado, se han retirado o no cumplen con ninguna de las condiciones anteriores.
Claserend	Determina la cantidad de materias perdidas por el estudiante.
Claseprom edio	Determina el promedio acumulado del estudiante.

Finalmente, se obtuvo el conjunto de datos UDENAR.ARFF con 26 atributos y 20329 registros, listo para aplicarle las técnicas de minería de datos, utilizando la herramienta TariyKDD, que permitan obtener los patrones de bajo rendimiento académico y /o deserción de los estudiantes de la Universidad de Nariño.

Tabla 4. Discretización del atributo Edad

Edad	Valor	No. Registros
Menores e iguales a 18	A	827
Mayores de 18 y menores que 22	B	3634
Mayores e iguales que 22 y Menores de 26	C	4856
Mayores e iguales que 26	D	11012

Tabla 5. Discretización del atributo Fecha de Ingreso

Fecha Ingreso	Valor	No. Registros
Antes de 1990	A	1022
Después o igual a 1990 a Menores de 1995	B	4852
Después o igual a 1995 a Menores de 2000	C	5978
Después o igual al 2000 y Menores de 2003	D	5046
Mayores o iguales de 2003	E	3431

Tabla 6. Discretización del atributo Clasepromedio

Clasepromedio	Valor	No.Registros
Menor a 2	A	2391
Mayor o igual a 2 hasta 3	B	2934
Mayor o igual a 3 hasta 3.5	C	5166
Mayor o igual a 3,5 hasta 4,0	D	6850
Mayor o igual a 4.0 hasta 5.0	E	2988

3.4 Etapa de Minería de Datos

La etapa de minería de datos es la más característica del proceso DCBD[8]. El objetivo de esta etapa es la búsqueda y descubrimiento de patrones insospechados y de interés utilizando tareas de descubrimiento tales como clasificación [7], clustering [3], patrones secuenciales [2] y asociacion [1] entre otras. Para el descubrimiento de patrones de deserción estudiantil y bajo rendimiento académico se utilizaron las tareas de Clasificación y Asociación. Para generar las reglas de clasificación su utilizó el algoritmo C4.5[10] y para las reglas de Asociación, el algoritmo EquipAsso[12][13], disponibles en la herramienta TariyKDD[11].

3.4.1 Reglas de Clasificación. Para predecir los perfiles de bajo rendimiento académico, el conjunto de datos UDENAR.ARFF se clasificó escogiendo como clase el atributo *Clasepromedio*. Este atributo indica el rendimiento académico del estudiante basado en el promedio acumulado de las notas hasta el semestre cursado.

Entre las reglas de clasificación más representativas están:

- Si el estrato socioeconómico es 2, el ponderado de exámenes de estado ICFES está entre 50 y 70, es del Sur de Nariño, está en primer semestre y pertenece a la facultad de Ciencias Humanas, entonces su rendimiento es Bajo. El 68% con estas características se clasifican de esta manera.
- Si la edad de ingreso es menor o igual a 18 años, el estrato socioeconómico es 2, género masculino, el ponderado ICFES está entre 50 y 70, vive con la familia, es del Sur de Nariño, está en primer semestre, está en la facultad de Ciencias Naturales y Matemáticas, entonces su rendimiento es Bajo. El 67% con estas características se clasifican de esta manera.
- Si la edad de ingreso es menor o igual a 18 años, proviene de un colegio privado, el calendario del colegio es septiembre a junio, género femenino, es del Sur de Nariño, está en primer semestre y pertenece a la facultad de Ciencias Naturales y Matemáticas, entonces su rendimiento es Bajo. El 70% con estas características se clasifican de esta manera.

Para predecir los perfiles de deserción estudiantil se escogió como clase el atributo *Clase_al*. Este atributo indica si el estudiante no se ha retirado, ha reingresado o se retiró definitivamente de la Universidad.

Entre las reglas de clasificación más representativas están:

- Más del 50% de los estudiantes retirados que pertenecen a la facultad de ingeniería, reingresan.
- Los estudiantes retirados que pertenecen a las facultades de Ciencias Naturales y Matemáticas y Ciencias Humanas no reingresan.

3.4.2 Reglas de Asociación. Entre las reglas de Asociación más representativas, que permiten identificar relaciones no explícitas entre los atributos del conjunto de datos UDENAR.ARFF que involucran bajo rendimiento y deserción están:

- El 95% de los estudiantes que tienen promedio bajo está en primer semestre.
  El 10% de todos los estudiantes son de primer semestre y tienen promedio bajo.

- El 84% de los estudiantes retirados son de estrato socioeconómico 2 y provienen de municipios del Sur de Nariño. El 2.5% de todos los estudiantes se han retirado, son de estrato 2 y provienen del Sur de Nariño.
- El 89 % de los estudiantes retirados son de primer semestre, tienen un ponderado ICFES entre 50 y 70 y proceden del Sur de Nariño. El 2.5% de todos los estudiantes se han retirado, son de primer semestre, tienen un ponderado ICFES entre 50 y 70 y provienen del Sur de Nariño.
- El 88% de los estudiantes retirados, tienen una edad de ingreso menor que 18 años provienen del Sur de Nariño. El 2.5% de todos los estudiantes se han retirado, tienen una edad de ingreso menor que 18 años y son del Sur de Nariño.
- El 86% de estudiantes retirados terminaron su bachillerato en colegios públicos, son de primer semestre y provienen del Sur de Nariño. El 2.5% de todos los estudiantes se han retirado, terminaron su

bachillerato en colegios públicos, son de primer semestre y provienen del Sur de Nariño.

3.5 Etapa de Interpretación y Evaluación de Resultados

De acuerdo a los resultados obtenidos, la mayoría de los estudiantes de primer semestre, provenientes de la zona sur del departamento de Nariño, de estratos socioeconómicos bajos y matriculados en algún programa de la facultad de Ciencias Naturales y Matemáticas o en la facultad de Ciencias Humanas, presentan un bajo rendimiento académico. Este perfil es similar al perfil de de la mayoría de estudiantes que se retiran. Por otra parte la mayoría de estudiantes que se retiran de estas dos facultades no reingresan, lo que no sucede en la facultad de Ingeniería, donde casi la mayoría de estudiantes retirados reingresan.

4. CONCLUSIONES Y RECOMENDACIONES

Se han presentado los resultados del proyecto de investigación cuyo objetivo era detectar patrones de bajo rendimiento académico y deserción estudiantil en la Universidad de Nariño, utilizando las tareas de minería de datos Clasificación y Asociación. Dentro de este proyecto, las fases de preprocesamiento y transformación de datos fueron las más costosas en tiempo, debido a la mala calidad de los datos de las bases de datos existentes.

En cuanto a los patrones obtenidos, la Universidad de Nariño debe tomar decisiones y proponer estrategias de seguimiento a estudiantes con estos perfiles con el fin de prevenir que caigan en bajo rendimiento y disminuir el grado de deserción que se presenta.

Con este proyecto, se demostró que TariyKDD es una herramienta fiable, que puede ser utilizada en cualquier proyecto de Minería de Datos y su distribución es libre.

Se recomienda la construcción de una bodega de datos que permita obtener datos de calidad para futuros proyectos de minería de datos que se realicen en la Universidad de Nariño.

REFERENCIAS

[1] Agrawal R., Srikant R. Fast Algorithms for Mining Association Rules, VLDB Conference, Santiago, Chile, 1994.

[2] Agrawal R., Srikant R.: Mining Sequential Patterns. In Proceedings of the 11th International Conference on Data Engineering, 1995.

[3] Berry, M., Linoff, G.: Data Mining Techniques for Marketing, Sales and Customer Support. Wiley Computer Publishing,1997.

[4] Chen M., Han J., Yu P. Data Mining: An Overview from Database Perspective. IEEE Transactions on Knowledge and Data Engineering, 1996.

[5] Fayyad, U., Piatetsky-Shapiro, G., Smyth P.: The KDD Process for Extracting Useful Knowledge from Volumes of Data. In Communications of the ACM, Vol. 39, No 11, November,1996.

[6] Han, J., Pei, J., Yin, Y., Mining Frequent Patterns without candidate Generation. Proc. ACM SIGMOD, Dallas, TX, 2000.

[7] Han, J., Kamber, M.: Data Mining Concepts and Techniques. Morgan Kaufmann Publishers, San Francisco, 2001.

[8] Hernandez, O.J., Ramirez, Q..M., Ferri, R.C.: Introducción a la Minería de Datos. Editorial Pearson Prentice Hall, Madrid, España ,2004.

[9] Imielinski, T., Mannila, H. A Database Perspective on Knowledge Discovery. Communications of the ACM, Vol 39, No. 11, November, 1996.

[10] Quinlan, J.R., C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, 1993.

[11] Timarán, P.R., Calderón, R.A., Ramírez, F.I., Guevara, F., Alvarado, J.C.: TariyKDD una Herramienta de Minería de Datos Débilmente Acoplada con un SGBD. In: Memorias de VII Jornadas Iberoamericanas de Ingeniería de Software e Ingeniería del Conocimiento, pp. 3—11. Editado por Escuela Superior del Litoral. Guayaquil, Ecuador, 2007.

[12] Timarán, P.R., Millán, M.,: EquipAsso: un Algoritmo para el Descubrimiento de Reglas de Asociación basado en Operadores Algebraicos. In Memorias de 4ª Conferencia Iberoamericana en Sistemas, Cibernética e Informática CISCI 2005, pp. 343—348. Orlando, Florida, USA, 2005.

[13] Timarán, P.R., Millán, M.: EquipAsso: an Algorithm based on New Relational Algebraic Operators for Association Rules Discovery. In proceedings of the Fourth IASTED International Conference on Computational Intelligence. ACTA Press, Calgary, Alberta, Canadá, 2005.

[14] Timarán, P.R.: Mate-tree: un Algoritmo para el Descubrimiento de Reglas de Clasificación basado en Operadores Algebraicos Relacionales. In Memorias de 6ª Conferencia Iberoamericana en Sistemas, Cibernética e Informática CISCI 2007, pp. 196—201. Orlando, Florida, USA, 2007.

[15] Witten, I. H., Frank, E: Data Mining practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann Publishers, San Francisco, California, USA, 2000.